Llumnix: Dynamic Scheduling for Large Language Model Serving

Year: 2024 | *Citations: 117* | *Authors: Biao Sun, Ziming Huang, Hanyu Zhao*

Abstract

Inference serving for large language models (LLMs) is the key to unleashing their potential in people's daily lives. However, efficient LLM serving remains challenging today because the requests are inherently heterogeneous and unpredictable in terms of resource and latency requirements, as a result of the diverse applications and the dynamic execution nature of LLMs. Existing systems are fundamentally limited in handling these characteristics and cause problems such as severe queuing delays, poor tail latencies, and SLO violations. We introduce Llumnix, an LLM serving system that reacts to such heterogeneous and unpredictable requests by runtime rescheduling across multiple model instances. Similar to context switching across CPU cores in modern operating systems, Llumnix reschedules requests to improve load balancing and isolation, mitigate resource fragmentation, and differentiate request priorities and SLOs. Llumnix implements the rescheduling with an efficient and scalable live migration mechanism for requests and their in-memory states, and exploits it in a dynamic scheduling policy that unifies the multiple rescheduling scenarios elegantly. Our evaluations show that Llumnix improves tail latencies by an order of magnitude, accelerates high-priority requests by up to 1.5x, and delivers up to 36% cost savings while achieving similar tail latencies, compared against state-of-the-art LLM serving systems. Llumnix is publicly available at https://github.com/AlibabaPAI/llumnix.